IN THE SPECIFICATION:

Please amend the paragraph beginning at **line 15 of page 8** to read as follows:

--Next, each utterance is encoded in terms of a set of 43 features that have the potential to be used during runtime to alter the course of the dialog. These features were either derived from prior encoding processes or from system 100. ~~these~~ features. The system 100 components from which ~~that~~ we extracted features ~~from~~ were the recognizer 120, the NLU unit 130, and the dialog manager 190, along with a representation of the discourse history. Because the contribution of different system 100 components to the problem of predicting NLU error are to be examined, a classifier that had access to all the features is trained and its performance is compared to classifiers that only had access to recognizer 120 features, to NLU unit 130 features, and to discourse contextual features. Below we describe features obtained from each module:

Recognizer Features

- recog, recog-numwords, asr-duration, stmf-flag, rg-modality, rg-grammar, tempo

NLU Unit Features

- a confidence measure for all of the possible tasks that the user could be trying to do
- salience-coverage, inconsistency, context-shift, top-task, nexttop-task, top-confidence, diff-confidence, confpertime, salpertime

Dialog Manager and Discourse History Features

- sys-LABEL, utt-id, prompt, reprompt, confirmation, sub-dial,
- discourse history: num-reprompts, num-confirms, num-subdials, reprompt%, confirmation%, subdialog%--

Please amend the paragraph beginning at **line 27 of page 9** to read as follows:

--One motivation for the *tempo* feature is that it has been found that users tend to slow down their communication (speech, movement, etc) when the system 100 has

misunderstood them. This strategy actually leads to more errors since the recognizer 120 is not trained on this type of communication. The *tempo* feature may also indicate hesitations, pauses or interruptions, which could also lead to recognizer 120 errors. On the other hand, additional multimodal input such as touchtone (DTMF) in combination with the user's communication, as encoded by the feature *dtmf-flag*, might increase the likelihood of understanding, since the touchtone input is unambiguous ~~it can~~ and will therefore function to constrain language understanding.--

Please amend the paragraph beginning at **line 29 of page 10** to read as follows:

--In addition, similar to the way we calculated the *tempo* feature, the *salience-coverage* and *top-confidence* features are normalized by dividing them by *asr-duration* to produce the *salpertime* and *confpertime* features. The for these NLU features is to make use of ~~in formation~~ information that the NLU unit 130 has a result of processing the output of recognizer 120 and the current discourse context. For example, fur utterances that follow the first utterance, the NLU unit 130 know what task it believes the user is trying to complete. The *context-shift* feature incorporates this knowledge of the discourse history, with the motivation that if it appears that the caller has changed his/her mind, then the NLU unit 130 may have misunderstood an utterance. –

Please amend the paragraph beginning at **line 20 of page 15** to read as follows:

--Fig. 5 is a flowchart of a possible automated task classification process using the natural language understanding monitoring system of the invention. The process ~~Process~~ begins ~~its~~ at step 5000 ~~in~~ and proceeds to step 5100 where an input communication is received by the recognizer 420. At step 5200, the recognizer 420 attempts to recognize portions of the user's input communication, including grammar fragments, meaningful words/phrases/symbols, morphemes, actions, gestures, or any other communication signal.--

Please amend the paragraph beginning at **line 3 of page 16** to read as follows:

--However, in step 5300, if the NLU monitor 180 determines that the task classification processor 440 cannot classify the user's request, in step 5500, the NLU

monitor 180 determines whether the probability of correctly understanding the user's input communication exists above a predetermined threshold. In this iteration, the NLU monitor 180 is using only the first exchange (exchange 1). The NLU monitor 180 sues the classification model stored in the dialog training database 165 to determine whether the probability of correctly understanding the user's input communication exceeds the predetermined threshold. If the NLU monitor 180 determines that the probability of correctly understanding the user's input communication does not exceed the threshold, then in step 5800 ~~for~~ the user is routed to a human for assistance. The process then goes to step 5900 and ends.--

Please amend the paragraph beginning at **line 28 of page 17** to read as follows:

--The output of each experiment is a classification model learned from the training data that is stored in the training database 165. The model is evaluated in several ways. First, multiple models are trained using different features sets extracted from different system 100 components in order to determine which feature sets are having the largest impact on performance. Second, for each feature set, the error rates of the learned classification models are estimated using ten-fold cross-validation, by training on a random 10.608 utterances and testing on a random 1,179 utterances 10 successive times. Third, precision, recall and the confusion matrix are recorded in the classifier, and then trained on all the features tested on a random held-out 20% test set. Fourth, for the classifier trained on all the features, the extent to which the error can be minimized on the error classes RMISMATCH and RPARTIAL-MATCH is examined by manipulating the rule-learning program's loss ratio parameter. Finally, the results of training other learners are compared on the same dataset with several of the feature sets. The overall accuracy results for detecting NLU errors using the rule-learning program are summarized ~~in Fig. 7~~ below (SE=Standard Error):~~.~~

Features Used	Accuracy	(SE)
BASELINE (majority class)	63.47%	
ALL	86.16%	(0.38)
NLU UNIT ONLY	84.80%	(0.38)

RECOGNIZER + DISCOURSE	80.97%	(0.26)
RECOGNIZER ONLY	78.89%	(0.27)
DISCOURSE ONLY	71.97%	(0.40)--